基于字典学习的跨媒体检索技术 *

戚玉丹，张化祥†，刘一鹤

(山东师范大学 信息科学与工程学院，济南 250358)

摘 要：在研究跨媒体信息检索时，对于不同模态数据的异构性提出了挑战，针对如何更好的克服异构问题以提高多模态数据之间的检索精度，提出了一种基于字典学习的新跨媒体检索技术。首先，通过字典学习方法学习两个不同模态数据之间的稀疏系数，然后，通过特征映射方案由两个不同的投影矩阵分别把它们投入共同的特征子空间，最后，通过标签对齐同一类来增强不同模态之间的相关性。实验结果表明，与传统的同构子空间学习方法相比，基于字典的算法分类性能优越，该实验方法在两个数据集上优于几种最先进的方法。

关键词：跨媒体检索；字典学习；稀疏表示；模态独立；特征映射

中图分类号：TP391

Cross-media retrieval technology based on dictionary learning

Qi Yudan, Zhang Huaxiang†, Liu Yihe

(*School of Information Science & Engineering Shandong Normal University*, *Jinan* 250358, *China*)

Abstract: In the study of cross-media retrieval, how to capture and correlate heterogeneous features originating from different modalities remains a challenge. To cope with the aforementioned problems, this paper presented a novel cross-modal retrieval framework based on coupled dictionary learning. Firstly, it obtained sparse coefficients from different modalities by imposing dictionary learning. Then, it projected the data samples from different modalities into a common feature space. Moreover, it leveraged label information to align the cross-modal data sample pairs in the common space so as to encourage the inherent correlation across the different modalities. Simulation experimental results show that the method based on dictionary learning algorithm has superior recognition performance in comparison with the methods based on traditional mid-level feature subspace, experiment results on two public datasets demonstrate that our method outperforms several state-of-the-art methods.

Key Words: cross-modal retrieval; dictionary learning; sparse representation; modality-dependent; feature mapping

0 引言

早期的数据检索多针对单模态数据，即查询和检索的数据属于相同模态。例如，给定一个文本查询，单模态的方法直接与网络上的文本原数据进行匹配，而不是相一致的图像。通常这些单模态的方法不能应用于跨媒体检索。跨媒体检索是多媒体检索中基于内容的一个新的研究领域，由于不同模态的数据之间存在着异构性难以实现直接互检。如何解决不同模态数据之间的异构问题，从而实现多媒体数据之间的互检成为跨媒体检索领域的一个重要研究问题。

近年来，针对跨媒体检索提出了许多新的方法，通过挖掘不同模态之间潜在的关系，实现跨模态数据之间的互检。具体来说，最具权威的典型相关性分析（canonical correlation analysis,CCA）[1]是一种经典的特征学习方法，该方法通过最大化两组特征之间的相关性，得到两种特征在子空间中的低维表达，并使其相关度最高。CCA 的提出为跨媒体检索的研究起到了很大的推动作用，它的扩展方法在跨媒体检索领域得到了广泛的应用。例如，Rasiwasia 等人[2]提出的方法中，从关联假设和抽象假设两个方面对跨媒体检索问题进行了整合。Hwang 等人[3]已经根据用户提供的注释顺序，对单词的相对重要性进行了建模，以提高跨模式检索的精度。Ballan 等人[4]使用核 CCA (kernel canonical correlation analysis,KCCA)开发交叉视图检索方法来建立图像和文本的关联性。除基于 CCA 的方法外，还有许多其他的跨媒体检索的方法。其中偏最小二乘法（partial least squares,PLS）[5]是一种新型的多元统计数据分析方法，它于 1983 年由伍德(Wold)和阿巴诺(Albano)等人首次提出，近年来，它在理论、方法和应用方面得到了迅速的发展。Chen 等人[6]将偏最小二乘法（PLS）应用于跨媒体检索，他们使用 PLS 来转换视

基金项目：国家自然科学基金资助项目（61373081）；山东省泰山学者项目

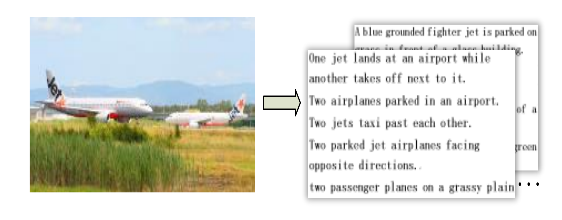
作者简介：戚玉丹（1991-），女，山东济宁人，硕士研究生，主要研究方向为异构媒体检索、云计算、机器学习；张化祥（1966-），男（通信作者），山东济宁人，院长，教授，博导，博士，主要研究方向为异构媒体检索、大数据分析技术、机器学习、模式识别（huaxzhang@163.com）；刘一鹤（1993-），女，山东泰安人，硕士研究生，主要研究方向为异构跨媒体检索、云计算、模式识别.

觉特征到文本特征空间中，学习一个语义来测量两种不同模式之间的相似性。除此之外，还有线性判别分析（LDA）和边界Fisher 分析(MFA)。Sharma 等人[7]将基于线性判别分析(LDA）和边界 Fisher 分析（MFA）的广义多视图分析扩展为广义多视图 LDA（GMLDA）和广义多视图 MFA（GMMFA）已应用于跨媒体检索。跨媒体哈希是通过将不同形式的数据嵌入到一个普通的低维汉明空间中进行跨媒体检索，将高维数据对象映射为简洁的哈希编码，使得相似的数据对象拥有相同或者相似的哈希码，进而通过测量二进制哈希码之间的相似度来获得原始数据之间的相似度，近年来引起了广泛的关注。例如，Yu 等人[8]提出了一种判别的哈希字典学习方法(DCDH)。另外，随着深度学习在计算机领域突破性的进展，一些深度学习方法也被用于跨媒体相似度检索模型，如卷积神经网络（convolutional neural network,CNN）、递归神经网络（recursive neural network,RNN）和自动编码（auto encoder）等。基于深度学习方法研究，Andrew 等人[9]提出了深度典型相关性分析（deep canonical correlation analysis,DCCA）。DCCA 学习不同模态数据之间的非线性投影,从而使得学习到的数据是高度线性相关的。Wang 等人[10]提出了基于有监督方法的多模态深层神经网络（MDCCN）。Jiang 等人[11]提出了基于深度学习的实时网络跨媒体检索方法，根据图像特征向量的贡献对它们中的元素进行排序，然后消除不必要的特性。即使这些存在的方法解决了跨媒体检索的问题，但是大多数存在的方法只专注于通过两个特征空间的距离来学习两种模态的相关性，从而忽略了不同的语义特征。另外，类标签信息也没有得到充分的利用。为充分学习在不同特征空间中的异构特征，稀疏字典学习日益受到广泛的关注。

1　简介

字典学习[12~16]旨在从训练数据中找到一组特殊的稀疏编码，这一组稀疏元素足以线性地表示这些原始数据的特征，从而用尽可能少的数据表示尽可能多的内容。因此，字典实质上是对于庞大数据集的降维，稀疏表示是用尽可能少的数据表示尽可能多的特征，以提高检索效率。由于这种表示是有效的，字典学习得到了广泛的应用。在本文中主要关注在图像与文本之间的多媒体检索(图 1)，使用图像搜索文本文档或者文本搜索图像（I2T 和 T2I）。其中图 1（a）给定一个飞机的图像，任务是找到与此图像相关的文本报告；（b）关于两个飞行员的文本文件，任务是找到关于他们相关的图片。

另一方面，本文将两个模态的检索任务分开来执行，即为模态独立方法。模态独立[17]不同于以前的方法学习一对投影，它学习两对映射将图像检索文本和文本检索图像从其原始特征空间投影到两个公共潜在子空间。因为如果两个任务同时学习，得到的公共子空间为 I2T 和 T2I 共同的最优子空间，通常对用于检索模态的语义理解并非最优的。例如，在图像检索文本中，通常认为图像语义空间中查询的准确表示比要检索的文本更为重要，若查询的语义被错误判断，则更难以检索相关文本。如果分开执行，图像检索文本时就可以把图像单独投影到它的语义空间，这时对这个图像的语义理解没有了文本的干扰则是最优的，理解了图像语义之后，对数据的检索更加准确，从而提高跨媒体检索的精度，通过实验证明模态独立相对其他算法也是有效的。



a) 用图像检索文本(I2T)



b) 用文本检索图像(T2I)

图 1　跨媒体检索任务

本文将字典学习与模态独立相结合，学习一种基于模态独立与字典学习新的跨媒体检索技术。首先通过字典学习将多模态数据转换为稀疏表示，并保证所生成的表示是均匀的；然后使用线性回归映射这些稀疏系数，将来自不同模态的数据生成的稀疏表示由两个不同的投影矩阵映射到两个公共语义空间中。图 2 描述了本文提出方法的框架。其中图 2（a）和（c）是两个线性回归操作，分别表示图像和文本特征空间到语义空间。因此，具有相同语义的多模态数据可以在公共潜在子空间中关联起来。图 2(b)是一种相关性分析操作，在公共空间中保持多模态数据的相互关联。将图 2（a）与（b）结合起来，学习对 I2T 的投影；同样的，对 T2I 学习一个不同的投影由图 2（b）和（c）共同优化。

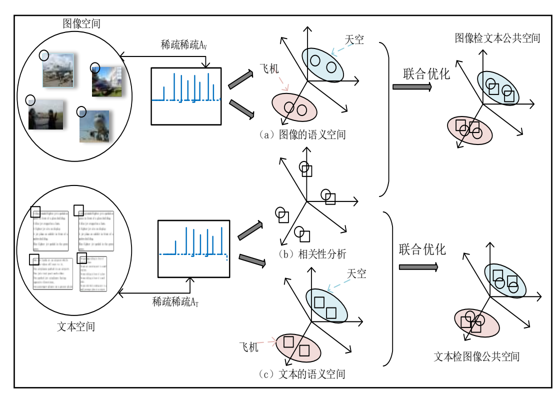


图 2　跨媒体检索的框架

2　相关工作

为了提高跨媒体检索效率，本文提出一种基于模态独立的

字典学习方法，其中字典学习模型是关键的技术，并将图像检索文本与文本检索图像的任务分开训练；最后确定目标函数，根据不同的参数设置来讨论它们的优化算法。

2.1 字典学习

设 $X=[x_1,x_2,...x_N]\in R^{N\times P}$ 是维度为 P，样本数为 N 的数据集。利用稀疏字典进行跨媒体检索中异构数据的处理。主要方法是通过字典重构，将原始数据之间的相关关系转换为稀疏系数之间的关系。其目标函数表示为

$$\min_{D,A}\|X-AD\|_F^2+\alpha\|A\|_1$$
$$s.t.\ \|d_i\|\le 1\ \ \forall i\in[1:K]\quad (1)$$

其中：$D=[d_1,d_2,...d_K]\in R^{K\times P}$ 是学习的字典；d_i 是字典中第 i 个原子；K 表示字典的大小； $A\in R^{N\times K}$ 是根据字典 D 得到的样本数据 X 的稀疏系数；$\|\|_F^2$ 表示 F 的二范数；$\|X-AD\|_F^2$ 目的是使字典 D 与稀疏系数 A 的线性组合尽可能地接近数据样本 X；另外，用 $\alpha\|A\|_1$ 来控制稀疏。

2.2 检索任务描述

综上对本文提出优化框架原理的概述，下面对本文检索的两个任务分别进行详细的描述。

2.2.1 图像检索文本

本节首先讨论跨媒体检索中图像检索相一致的文本。其中 I2T 的线性回归术语是一个从图像空间到语义空间的回归操作。

假设本文定义 $X_V=[v_1,v_2,...v_n]\in R^{n\times p}$ 为维度 p，个数为 n 的图像数据集； $X_T=[t_1,t_2,...t_n]\in R^{n\times q}$ 为维度为 q，个数为 n 的文本数据集；$D_V\in R^{k\times p}$ 是学习图像的字典，$D_T\in R^{k\times p}$ 是学习文本的字典；$A_V\in R^{n\times k}$ 是图像的稀疏系数，$A_T\in R^{n\times k}$ 是文本的稀疏系数，其中图像的稀疏稀疏 A_V 和文本的稀疏系数 A_T 依赖于学习的字典得出。$Y^{(i)}=[y_1,y_2...y_n]\in R^{m\times c}$ 是关键词矩阵，即公共语义子空间。其中图像的投影矩阵是 $W_{V1}\in R^{k\times c}$，$W_{T1}\in R^{k\times c}$ 是文本的投影矩阵。字典学习的目的是分别学习图像和文本的两个投影矩阵，利用投影矩阵将两个模态的稀疏表示 A_V 和 A_T 投影到一个共同的特征空间中。描述的框架表述如下：

$$\min_{D_V,D_T,A_V,A_T,W_{V1},W_{T1}}\|X_V-A_VD_V\|_F^2+\|X_T-A_TD_T\|_F^2+\|A_VW_{V1}-Y^{(i)}\|_F^2+$$
$$\alpha_1(\|A_V\|_1+\|A_T\|_1)+\|A_VW_{V1}-A_TW_{T1}\|_F^2+\alpha_2\|W_{V1}\|_F^2+\alpha_3\|W_{T1}\|_F^2\quad (2)$$

其中：$\|X_V-A_VD_V\|_F^2$ 为通过字典学习图像的稀疏系数 A_V；$\|X_T-A_TD_T\|_F^2$ 为通过字典学习文本的稀疏系数 A_T；$\|A_VW_{V1}-Y^{(i)}\|_F^2$ 作为线性回归项，通过投影矩阵 W_{V1} 将稀疏系数矩阵投影到语义空间，使得具有相同语义的多媒体数据聚集在一起。其中参数 $0\le\alpha\le1$，为均衡参数。$\|A_V\|_1$ 和 $\|A_T\|_1$ 用来控制稀疏，$\|W_{V1}\|_F^2$ 和 $\|W_{T1}\|_F^2$ 用来控制投影矩阵 W_{V1} 和 W_{T1} 复杂度避免过拟合。$\|A_VW_{V1}-A_TW_{T1}\|_F^2$ 为相关分析项，目的是使同一类的数据更相近，增强不同模态之间的相关性。本文的模型中，不同模态的数据相关性得以表示。

2.2.2 文本检索图像

本节讨论跨媒体检索中文本检索相一致的图像。T2I 的线性回归术语是一个从文本空间到语义空间的回归操作，与图像检索类似。

定义 $X_V=[v_1,v_2,...v_n]\in R^{n\times p}$ 为维度 p，个数为 n 的图像数据集；$X_T=[t_1,t_2,...t_n]\in R^{n\times q}$ 为维度为 q，个数为 n 的文本数据集；$A_V\in R^{n\times k}$ 是图像的稀疏系数，$A_T\in R^{n\times k}$ 是文本的稀疏系数；$D_V\in R^{k\times p}$ 是学习图像的字典，$D_T\in R^{k\times p}$ 是学习文本的字典；$Y^{(t)}=[y_1,y_2...y_n]\in R^{n\times c}$ 是公共语义子空间，与 $Y^{(i)}$ 可以近似的看做一个公共语义子空间。这里设两个与图形检索文本不同的投影矩阵 $W_{V2}\in R^{k\times c}$ 和 $W_{T2}\in R^{k\times c}$，描述的框架表述如下：

$$\min_{D_V,D_T,A_V,A_T,W_{V2},W_{T2}}\|X_T-A_TD_T\|_F^2+\|X_V-A_VD_V\|_F^2+\|A_TW_{T2}-Y^{(t)}\|_F^2+$$
$$\alpha_1(\|A_V\|_1+\|A_T\|_1)+\|A_VW_{V2}-A_TW_{T2}\|_F^2+\alpha_2\|W_{V2}\|_F^2+\alpha_3\|W_{T2}\|_F^2\quad (3)$$

与图像检索文本原理相同，其中 $\|A_TW_{T2}-Y^{(t)}\|_F^2$ 为通过投影矩阵 W_{T2} 将稀疏系数矩阵投影到关键词子空间，使得具有相同语义的多媒体数据聚集在一起； $\|W_{V2}\|_F^2$ 和 $\|W_{T2}\|_F^2$ 控制其复杂度避免过拟合；$\|A_VW_{V2}-A_TW_{T2}\|_F^2$ 作为相关分析项使同一类的数据更相近，提高它们的相关性。同样的，在本文模型中，不同模态的数据相关性被表示。

3 优化

I2T 和 T2I 的优化问题是两个矩阵的无约束优化问题。因此，式（2）和（3）是非凸优化问题，并有许多局部最优解。为解决这个问题，设计一个算法来寻找固定点。可以注意到，当固定其他两项时，式（2）对另一项是凸面的。相似的，式（3）在固定另外两个的情况下，也可以是凸面的。分别通过固定 D_V（D_T）、A_V (A_T)或者 W_{V1}（W_{T2}）中的其中两个，用迭代更新来完成对另一个的最小化。具体优化策略如下：

首先，更新字典 D_V，固定稀疏系数 A_V 和投影矩阵 W_{V1}，如下：

$$\min_{D_V} \|X_V - A_V D_V\|_F^2 \quad s.t.\ \|d_i\| \le 1 \quad \forall i \in [1:K] \quad (4)$$

这是一个二次约束的二次规划问题（QCQP），求解可以通过拉格朗日对偶技术得到[20]。

同理，对于字典 D_T 的求解相似,可以由下式得出：

$$\min_{D_T} \|X_T - A_T D_T\|_F^2 \quad s.t.\ \|d_i\| \le 1 \quad \forall i \in [1:K] \quad (5)$$

然后，在字典 D_V 和投影矩阵 W_{V1} 不变的情况下来求解稀疏系数。由（2）可得

$$\min_{A_V} \|X_V - A_V D_V\|_F^2 + \|A_V W_{V1} - Y\|_F^2 + \alpha_1 \|A_V\|_1 + \|A_V W_{V1} - A_T W_{T1}\| \quad (6)$$

通过分析，求偏导可得

$$A_V = (X_V D_V^T + Y W_{V1}^T + A_T W_{T1} W_{V1}^T)^{-1} (D_V D_V^T + \alpha_1 E + 2 W_{V1} W_{V1}^T)^{-1}$$

同理可得

$$A_T = (X_T D_T^T - A_V W_{V1} W_{T1}^T)^{-1} (D_T D_T^T + \alpha_1 E - W_{T1} W_{T1}^T)^{-1}$$

最后，更新投影矩阵 W_{V1}，固定字典 D_V 和稀疏系数 A_V。分析由式（2）可得

$$\min_{W_{V1}} \|A_V W_{V1} - Y\|_F^2 + \|A_V W_{V1} - A_T W_{T1}\|_F^2 + \alpha_2 \|W_{V1}\|_F^2 \quad (7)$$

同理，可求解得

$$W_{V1} = (2 A_V^T A_V + \alpha_2 E)^{-1} (A_V^T Y + A_V^T A_T W_{T1})^{-1}$$

$$W_{T1} = A_V^T A_V W_V (A_V^T A_T + \alpha_3 E)^{-1}$$

综上所述，本文设计的目标函数在各部分均为凸函数，因此有最优解。为了获得最终结果，需要不断地重复上述步骤，直到最终收敛。本文在以下的算法中总结了此过程。相似的方法可以应用到文本检索图像。

算法　I2T 算法描述

算法 1I2T 的交替迭代优化过程

输入：图像的特征矩阵 X_V,，文本的特征矩阵 X_T，以及图像和文本相一致的语义 Y。

1 初始化字典 D_V 、D_T 和稀疏系数 A_V 、A_T 靠 FDDL[22]，设 W_{V1} 、W_{T1} 为单位矩阵。

2 如果不收敛则继续执行。

3 更新字典 D_V 、 D_T. 由式（4）（5），固定稀疏系数 A_V、A_T 和投影矩阵 W_{V1}、W_{T1}。

4 更新稀疏系数 A_V、A_T. 由式（6），固定字典 D_V、D_T 和投影矩阵 W_{V1}、W_{T1}。

5：更新投影矩阵 W_{V1}、W_{T1}. 由式（7），固定字典 D_V、D_V 和系数 A_V、A_T。

6：直到收敛为止。

输出：字典 D_V、D_T 和投影矩阵 W_{V1}、W_{T1}。

4 实验

为验证本文提出的跨媒体检索性能，进行了以下实验：首先阐述实验设置和本文采用的评估指标，然后将本文提出的方法与其他几种模型进行比较。

4.1 实验设计

本文在两个公共图像-文本数据集上对该方法进行评估。Wikipedia 文本图像数据集[17]和 Pascal Sentence 的数据集[17]。实验针对两个检索任务进行的:a)图像数据库中的文本查询；b)文本数据库中的图像查询。

Wikipedia 数据集：数据集包含有 10 个类的 2 866 个图像—文本对，随机地将数据集分为 2 173 个训练集和 693 个测试集。

Pascal Sentence 数据集：数据集包含了 1 000 个图像—文本对，由 20 个语义类别的标签标注(每个类别有 50 对)，对于每一类，随机选择 30 个图像—文本对作为训练集，其余的作为测试集。

对于两个数据集，每个图像—文本对的真实标签用来构造语义向量(用于 Wikipedia 数据集的 10 维，用于 Pascal Sentence 数据集的 20 个维度)被用于语义表示。具体地，本文利用了 4 096 维 CNN 视觉特征表示图像和由文献[17]所公开提供的的 100 维 LDA 来表示文本。

在本文中，使用归一化相关的系数（NC）来测量变换子空间中不同媒体对象的特征之间的相似度，通过召回率（PR）曲线和平均精度均值（mAP）来评估检索的性能。mAP 是每个查询的平均精度(AP)的平均值。分别的，定义平均精度为 $AP = \frac{1}{T}\sum_{r=1}^{N} P(r)\delta(r)$，其中：T 是属于同一类别的检索数据的数量；P(r)表示第 r 个检索数据的精度。如果第 r 个检索的数据与查询具有相同的标号，则 $\delta(r) = 1$，否则 $\delta(r) = 0$。在实验中，设置 N=50。查询所有的平均精度 AP 的值以获得平均精度的平均值 mAP，其中 mAP 的值越大，算法的准确性越高。

4.2 性能比较

为了客观地评价本文提出的方法，将本文所提出的方法与其他几种主要的算法进行比较。其中包括典型相关性分析 CCA

算法[2]、深度典型相关性分析 DCCA[9]、语义匹配 SM 算法[2]、语义关联匹配 SCM 算法[2]、三视图 CCA(TV CCA)[25]、广义多视角线性判别分析（GMLDA）[7]、广义多视图边缘 Fisher 分析（GMMFA）[7],以及模态独立的跨媒体检索（MDCR）[17]。在本文的实验中，所有的比较方法都使用相同的特性和训练集进行比较。

$\alpha_1=0.1$、$\alpha_2=0.5$ 、$\alpha_3=0.5$，用于优化 I2T 和 T2I。比较结果显示在表 1 中，可以看出本文提出的方法平均精度均值 mAP 从 1.9%平均改善至 18.4%。图像查询文本任务和文本查询图像任务的精确范围曲线显示在图 3 中，范围是检索到的顶级数据的数量。可以观察到，本文方法有更好的结果，它优于几种最先进的方法。

表 2 Wikipedia 数据集的跨媒体检索性能比较

方法	平均精度均值（mAP）		
	图像检索文本	文本检索图像	平均值
CCA	0.226	0.246	0.236
DCCA	0.309	0.288	0.298
SM	0.403	0.357	0.380
SCM	0.351	0.324	0.337
T-VCCA	0.310	0.316	0.313
GMLDA	0.372	0.322	0.347
GMMFA	0.371	0.322	0.346
MDCR	0.420	0.382	0.401
Proposed	**0.438**	**0.401**	**0.420**

表 3 Pascal Sentence 数据集的跨媒体检索性能比较

方法	平均精度均值（mAP）		
	图像检索文本	文本检索图像	平均值
CCA	0.261	0.356	0.309
DCCA	0.322	0.366	0.344
SM	0.426	0.467	0.446
SCM	0.369	0.375	0.372
T-VCCA	0.337	0.439	0.388
GMLDA	0.456	0.448	0.462
GMMFA	0.455	0.447	0.451
MDCR	0.448	0.475	0.462
Proposed	**0.483**	**0.490**	**0.486**

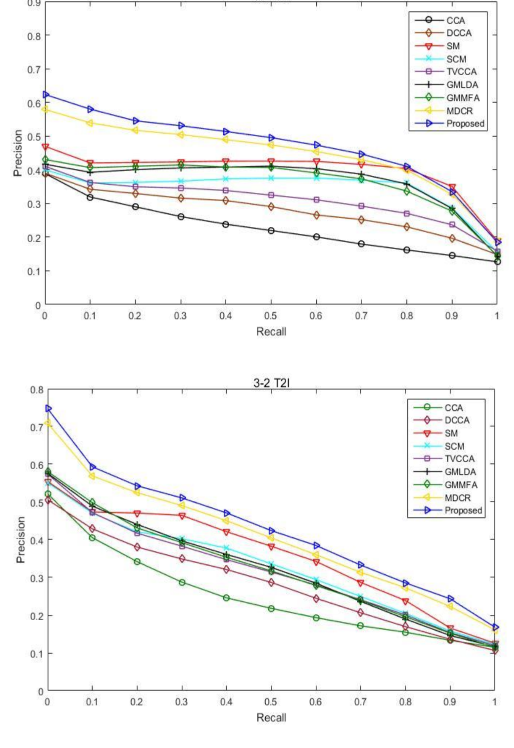


图 3 Wikipedia 数据集上召回率比较

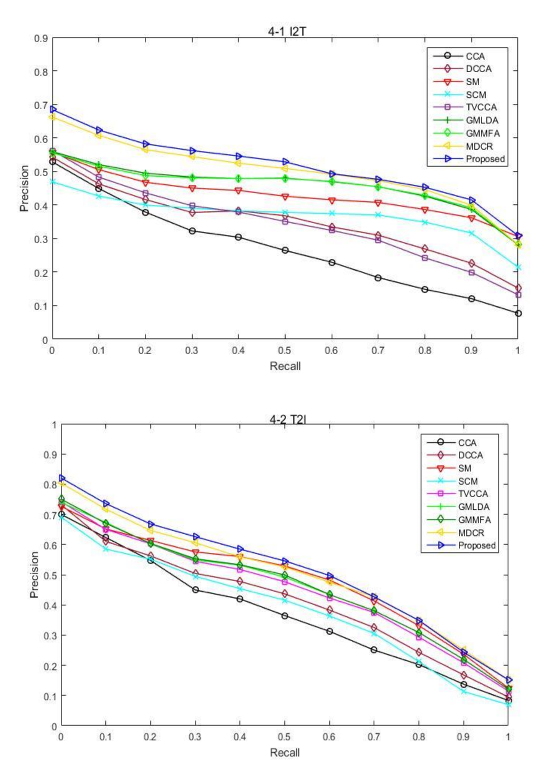


图 4 Pascal Sentence 数据集上召回率比较

在实验中，μ 是交替更新过程中的步长，ε 是收敛的条件，因此，设它们的范围在 0~1 间。它们的值越小，则交替更新的结果越准确。在测试集上进行实验的参数是根据训练集的交叉验证结果确定，而不是任意选择。

在 Wikipedia 数据集上，在测试了不同的参数设置后，首先确定了$\mu=0.02$，$\varepsilon=10^{-2}$。为了进一步验证实验效率，选用 4 096 维 CNN 的图像特征和 100 维 LDA 的文本特征。实验中设置

在 Pascal Sentence 数据集上，设置$\mu=0.02$，ε=10-4，$\alpha_1=0.01$、$\alpha_2=0.5$ 、$\alpha_3=0.5$，用于优化 I2T 和 T2I。比较结果显示在表 2 中，本文提出的方法平均精度 mAP 平均改善从 2.4%至 17.7%。图像查询文本任务和文本查询图像任务的精确范围曲线显在图 4 中，在实验中，本文方法对两个任务都得到了更好的结果。

5 结束语

本文设计了一个有效的跨媒体检索模型，通过字典学习生成稀疏系数，并将不同形式的数据投射到公共子空间，利用标签对齐方式增强不同模式之间的相关性，在这个空间中可以很好地发挥模式之间的内在联系；另外，本文将图像搜索文本与文本搜索图像分开来训练，分别来学习两对投影，充分发挥了它们各自的特征优势。在 Wikipedia 数据集和 Pascal Sentence 两个数据集上，大量的实验证明，提出的方法不仅提高了多模态之间的检索效率，而且对于单模态数据的识别也是有效的，为字典学习扩展了稀疏表示，对于求解最小化问题提出了有效的迭代算法。实验结果表明，本文提出的方法是有效的。

参考文献：

[1] Hardoon D, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: an overview with application to learning methods [J]. Neural Comput, 2004, 16 (12): 2639–2664

[2] Rasiwasia N, Pereira J, Coviello E, et al. A new approach to cross-modal multimedia retrieval [C]// Proc of the 18th ACM International Conference on Multimedia. 2010: 251-260.

[3] Hwang S, Grauman K. Learning the relative importance of objects from tagged images for retrieval and cross-modal search [J]. International Journal of Computer Vision, 2012, 100: 134-153.

[4] Ballan L, Uricchio T, Seidenari L, et al. A cross-media model for automatic image annotation [C]// Proc of International Conference on Multimedia Retrieval. Glasgow, United Kingdom: ACM Press. 2014: 73.

[5] Rosipal R, Krämer N. Overview and recent advances in partial least squares [C]// Proc of International Conference on Subspace, Latent Structure and Feature Selection. Bohinj, Slovenia: Springer-Verlag, . 2005: 34-51.

[6] Chen Y, Wang L, Wang W, et al. Continuum regression for cross-modal multimedia retrieval [C]// Proc of IEEE International Conference on Image Processing. 2013: 1949-1952.

[7] Sharma A, Kumar A, Daume H, et al: Generalized multiview analysis: a discriminative latent space [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2012: 2160–2167.

[8] Yu Z, Wu F, Yang Y, et al. Discriminative coupled dictionary hashing for fast cross-media retrieval [C]// Proc of International ACM SIGIR Conference on Research & Development in Information Retrieval. Gold Coast, Queensland, ACM Press, 2014: 395-404.

[9] Andrew G, Arora R, Bilmes JA, et al. Deep canonical correlation analysis [C]// Proc of International Conference on Machine Learning. 2013: 1247-1255.

[10] Wang W, Yang X, Ooi B C, et al. Effective deep learningbased multi-modal retrieval [J]. The VLDB Journal, 2016, 25 (1): 79–101.

[11] Jiang B, Yang J, Lv Z, et al. Internet cross-media retrieval based on deep learning [J]. Journal of Visual Communication & Image Representation, 2017, 48: 356-366.

[12] Pan Q, Liang Y, Zhang L, et al. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S. l.] : IEEE Computer Society, 2012: 2216-2223.

[13] Zhuang Y, Wang Y F, Wu F, et al: Supervised coupled dictionary learning with group structures for multi-modal retrieva [C]// Proc of the 27th AAAI Conference on Artificial Intelligence. 2013: 1070-1076.

[14] Huang Dean, Wang Y C F. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition [C]// Proc of IEEE International Conference on Computer Vision. 2013: 2496-2503.

[15] Xu X, Yang Y, Shimada A, et al. Semi-supervised coupled dictionary learning for cross-modal retrieval in Internet images and texts [C]// Proc of ACM International Conference on Multimedia. 2015: 847-850.

[16] Xu X S. Dictionary Learning Based Hashing for Cross-Modal Retrieval [C]// Proc of ACM on Multimedia Conference. 2016: 177-181.

[17] Wei Y, Zhao Y, Zhu Z, et al. Modality-dependent cross-media retrieval [J]. ACM Trans on Intelligent Systems and Technology, 2016, 17 (4): 57.

[18] Wang Kaiye, He Ran, Wang Wei, et al. Learning coupled feature spaces for cross-modal matching [C]// Proc of IEEE International Conference on Computer Vision. 2013: 2088-2095.

[19] Putthividhy D, Attias H T, Nagarajan S S. Topic regression multi-modal Latent Dirichlet Allocation for image annotation [C]// Proc of IEEE International Conference on Computer Vision and Pattern Recognition. 2010: 3408-3415.

[20] Schölkopf B, Platt J, Hofmann T. Efficient sparse coding algorithms [C]// Advances in Neural Information Processing Systems. 2006: 801-808.

[21] Wu Fei, Han Yahong, Liu Xiang, et al. The heterogeneous feature selection with structural sparsity for multimedia annotation and hashing: a survey [J]. International Journal of Multimedia Information Retrieval, 2012, 1 (1): 3-15.

[22] Yang M, Zhang D, Feng X. Fisher discrimination dictionary learning for sparse representation [C]// Proc of IEEE International Conference on Computer Vision. 2011: 543-550.

[23] Rasiwasia N, Mahajan D, Mahadevan V, et al. Cluster canonical correlation analysis [C]// Proc of the 17th International Conference on Artificial Intelligence and Statistics. 2014: 823-831.

[24] Wang Yanfei, Wu Fei, Song Jun, et al. Multi-modal mutual topic reinforce modeling for cross-media retrievalss [C]// Proc of the 22nd ACM International Conference on Multimedia. 2014: 307-316.

[25] Gong Y, Ke Q, Isard M, et al. A multi-view embedding space for modeling Internet images, tags, and their semantics [J]. International Journal of Computer Vision, 2014, 106 (2): 210-233.

[26] Cao Y, Long M, Wang J, et al. Deep visual-semantic hashing for cross-modal retrieval [C]// Proc of ACM SIGKDD International Conference on

Knowledge Discovery and Data Mining. San Francisco, California: ACM Press, 2016: 1445-1454.

[27] Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning [C]// Proc of International Conference on Machine Learning. 2011: 689-696.

[28] Shang X, Zhang H, Chua T S. Deep learning generic features for cross-media retrieval [M]// MultiMedia Modeling. Miami, FL: Springer International Publishing, 2016: 264-275.

[29] Feng F, Wang X, Li R. Cross-modal Retrieval with Correspondence Autoencoder [C]// Proc of International Conference on Multimedia. Orlando, Florida: ACM Press, 2014: 7-16.

[30] Cao Yue, Long Mingsheng, Wang Jiamin, et al. Correlation hashing network for efficient cross-modal retrieval [C]// Proc of IEEE International Conference on Computer Vision and Pattern Recognition. 2016.